EECS 598: Statistical Learning Theory, Winter 2014 **Topic 6**

Sieve Estimators: Consistency and Rates of Convergence

Lecturer: Clayton Scott *Scribe: Julian Katz-Samuels, Brandon Oselio, Pin-Yu Chen*

1 Introduction

We can decompose the excess risk of a discrimination ruleas follows:

$$R(\widehat{h}_n) - R^* = \underbrace{R(\widehat{h}_n) - R^*_{\mathcal{H}}}_{\text{estimation error}} + \underbrace{R^*_{\mathcal{H}} - R^*}_{\text{approximation error}}$$

The first term is the *estimation error* and measures the performance of the discrimination rule with respect to the best hypothesis in $\mathcal{H}$. In previous lectures, we studied performance guarantees for this quantity when $\widehat{h}_n$ is ERM. Now, we also consider the approximation error, which captures how well a class of hypotheses $\{\mathcal{H}_k\}$ approximates the Bayes decision boundary. For example, consider the histogram classifiers in Figure 1 and Figure 2, respectively. As the grid becomes more fine-grained, the class of hypotheses approximates the Bayes decision boundary with increasing accuracy. The examination of the approximation error will lead to the design of sieve estimators that perform ERM over $\mathcal{H}_k$ where $k = k(n)$ grows with n at an appropriate rate such that both the approximation error and the estimation error converge to 0. Note that while the estimation error is random (because it depends on the sample), the approximation error is not random.

2 Approximation Error

The following assumption will be adopted to establish universal consistency.

Definition 1. *A sequence of sets of classifiers $\mathcal{H}_1, \mathcal{H}_2, \ldots$ is said to have the* universal approximation property (UAP) *if for all P_{XY},*

$$R^*_{\mathcal{H}_k} \to R^*$$

as $k \to \infty$.

Observe that $R^*_{\mathcal{H}_k}$ is not random; it does not depend on the training data. Also, it depends on P_{XY} because it is an expectation. The following theorem gives a sufficient condition for a sequence of classifiers to have the UAP.

Theorem 1. *Suppose $\mathcal{H}_k$ consists of classifiers that are piecewise constant on a parititon of $\mathcal{X}$ into cells $A_{k,1}, A_{k,2}, \ldots$. If $\sup_{j \geq 1} \operatorname{diam}(A_{k,j}) \to 0$ as $k \to \infty$, then $\{\mathcal{H}_k\}$ has the UAP.*

Proof. See the proof of Theorem 6.1 in [1]. □

Example. Suppose $\mathcal{X} = [0,1]^d$. Consider $\mathcal{H}_k = \left\{\text{histogram classifiers based on cells of sidelength } \frac{1}{k}\right\}$. Then, $\operatorname{diam}(A_{k,j}) = \frac{\sqrt{d}}{k}$ for $\forall j$. In particular,

$$\begin{aligned} \sup_{j \leq 1} \operatorname{diam}(A_{k,j}) &= \frac{\sqrt{d}}{k} \\ &\to 0 \end{aligned}$$

as $k \to \infty$. By the above theorem, $\{H_k\}$ has the UAP.

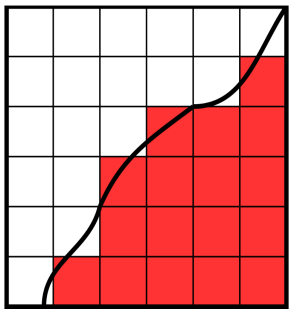

Figure 1: Histogram classifier based on 6×6 grid

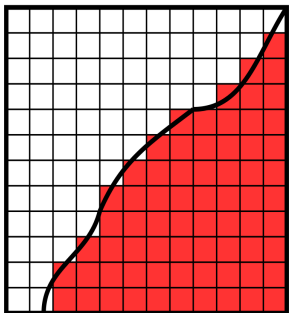

Figure 2: Histogram classifier based on 12×12 grid

3 Convergence of Random Variables

This section offers a brief review of convergence in probability and almost surely, and gives some useful results for proving both kinds of convergence.

Definition 2. *Let $Z_1, Z_2, \ldots$ be a sequence of random variables in $\mathbb{R}$. We say that Z_n converges to Z in probability, and write $Z_n \xrightarrow{i.p.} Z$, if $\forall \epsilon > 0$,*

$$\lim_{n\to\infty} \Pr(|Z_n - Z| \geq \epsilon) = 0.$$

We say Z_n converges to Z almost surely (with probability one), and write $Z_n \xrightarrow{a.s.} Z$, if

$$\Pr(\{\omega : \lim_{n\to\infty} Z_n(\omega) = Z(\omega)\}) = 1.$$

Example. In analysis of discrimination rules, we consider $\Omega = \{\omega = ((X_1, Y_1), (X_2, Y_2), \ldots) \in (\mathcal{X} \times \mathcal{Y})^\infty\}$ where $(X_i, Y_i) \overset{i.i.d.}{\sim} P_{XY}$. To study convergence of the risk to the Bayes risk, we take $Z_n = R(\widehat{h}_n)$ where $\widehat{h}_n$ is a classifier based on the first n entries of ω, namely $(X_1, Y_1), \ldots, (X_n, Y_n)$, and $Z = R^*$.

Now, we consider some useful lemmas for proving convergence in probability and almost sure convergence. We begin with useful results for convergence in probability.

Lemma 1. *If $Z_n \xrightarrow{a.s.} Z$, then $Z_n \xrightarrow{i.p.} Z$.*

Proof. This was proved in EECS 501. □

Lemma 2. *If $\mathbb{E}\{|Z_n - Z|\} \to 0$ as $n \to \infty$, then $Z_n \xrightarrow{i.p.} Z$.*

Proof. By Markov's inequality

$$\begin{aligned}\Pr(|Z_n - Z| \geq \epsilon) &\leq \frac{\mathbb{E}\{|Z_n - Z|\}}{\epsilon} \\ &\to 0.\end{aligned}$$

□

Now, we turn our attention to almost sure convergence. We begin with the Borel-Cantelli Lemma, which yields a useful corollary.

Lemma 3 (Borel-Cantelli Lemma)**.** *Consider a probability space* (Ω, A, P)*. Let* $\{A_n\}_{n\geq 1}$ *be a sequence of events. If*

$$\sum_{n\geq 1} P(A_n) < \infty$$

then

$$P(\limsup A_n) = 0$$

where

$$\begin{aligned}\limsup A_n &:= \{\omega \in \Omega : \forall N\, \exists n \geq N \text{ such that } \omega \in A_n\} \\ &= \{\omega \text{ that occur infinitely often in the sequence of events}\} \\ &= \bigcap_{N\geq 1} \bigcup_{n\geq N} A_n.\end{aligned}$$

Proof. Observe that $P(\limsup A_n) = P(\lim_{N\to\infty} \bigcup_{n\geq N} A_n)$. Let $B_N = \bigcup_{n\geq N} A_n$. Observe that $B_1 \supset B_2 \supset B_3 \supset \cdots$, i.e., $\{B_N\}$ is a decreasing sequence of events. Then, by continuity of P, we have

$$\begin{aligned}P(\lim_{N\to\infty} B_N) &= \lim_{N\to\infty} P(B_N) \\ &= \lim_{N\to\infty} P(\bigcup_{n\geq N} A_n) \\ &\leq \lim_{N\to\infty} \sum_{n\geq N} P(A_n) \\ &= 0\end{aligned}$$

where the inequality in the third line follows from union bound and the last equality follows from the hypothesis that $\sum_{n\geq 1} P(A_n) < \infty$. □

The following corollary of the Borel-Cantelli lemma is very useful for showing almost sure convergence.

Corollary 1. *If for all* $\epsilon > 0$*, we have*

$$\sum_{n\geq 1} \Pr(|Z_n - Z| \geq \epsilon) < \infty$$

then $Z_n \xrightarrow{a.s.} Z$*.*

Proof. Define the event $A^\epsilon = \{\omega \in \Omega : \forall N\, \exists n \geq N$ such that $|Z_n(w) - Z(w)| \geq \epsilon\}$. Let $A_n^\epsilon = \{\omega \in \Omega : |Z_n(\omega) - Z(\omega)| \geq \epsilon\}$. Observe that $\limsup A_n^\epsilon = A^\epsilon$. By the hypothesis, we may apply the Borel-Cantelli lemma to obtain $\Pr(A^\epsilon) = 0$.

Now define $A = \{\omega \in \Omega : Z_n(\omega) \not\to Z\} = \{\omega \in \Omega : \exists \epsilon > 0\, \forall N\, \exists n \geq N$ such that $|Z_n(\omega) - Z(\omega)| \geq \epsilon\}$. In words, this means: for some ϵ, no matter how far along you go in the sequence, you can find some n such that $|Z_n(\omega) - Z(\omega)| \geq \epsilon$. Now, let $\{\epsilon_j\}$ be a strictly decreasing sequence converging to 0. Observe that $A \subset \bigcup_{j=1}^\infty A^{\epsilon_j}$. To see this, take some $\omega \in A$. Then, for some $\epsilon > 0\ \forall N \exists n \geq N$ such that $|Z_n(\omega) - Z(\omega)| \geq \epsilon$. As $j \to \infty$, eventually there is some j' such that $\epsilon_{j'} \leq \epsilon$, so that $\omega \in A^{\epsilon_{j'}}$. Therefore,

$$\begin{aligned}\Pr(A) &\leq \Pr(\bigcup_{j=1}^\infty A_j^\epsilon) \\ &\leq \sum_{j=1}^\infty \Pr(A^{\epsilon_j}) \\ &= 0.\end{aligned}$$

This concludes the proof. □

4 Sieve Estimators

Let $\{\mathcal{H}_k\}$ be a sequence of sets of classifiers with the uniform approximation property (UAP). Assume that we have a uniform deviation bound for $\mathcal{H}_k$ (e.g., if $\mathcal{H}_k$ is finite or a VC class). We denote by $\widehat{h}_{n,k}$ the classifier we obtain from empirical risk minimization (ERM) over $\mathcal{H}_k$:

$$\widehat{h}_{n,k} = \arg\min_{h \in \mathcal{H}_k} \frac{1}{n} \sum_i \mathbf{1}_{\{h(X_i) \neq Y_i\}} \ .$$

Let $k(n)$ be an integer-valued sequence and define $\widehat{h}_n = \widehat{h}_{n,k(n)}$. Since we have the UAP, it should be clear that the approximation error goes to 0 as long as $k(n) \to \infty$ with n. We wish to restrict the rate of growth of $k(n)$ appropriately so that the estimation error also goes to zero, in probability or almost surely. This is called a *sieve estimator*, whose name is generally attributed to Grenander [3]. Formally, we need that

$$R(\widehat{h}_n) - R^*_{\mathcal{H}_{k(n)}} \longrightarrow 0, \text{ i.p/a.s.}$$

Let's apply some previously developed theory to this problem. Consider the case where $|\mathcal{H}_k| < \infty$. We have seen that

$$\Pr\left(R(\widehat{h}_n) - R^*_{\mathcal{H}_{k(n)}} \geq \epsilon\right) \leq 2\left|\mathcal{H}_{k(n)}\right| e^{-n\epsilon^2/2} \ . \tag{1}$$

We need to make sure that $\left|\mathcal{H}_{k(n)}\right|$ does not dominate the exponential term, so that the sum in Corollary 1 converges. Rewrite the right-hand side of Eqn. (1) as

$$\left|\mathcal{H}_{k(n)}\right| e^{-n\epsilon^2/2} = e^{\ln\left|\mathcal{H}_{k(n)}\right| - n\epsilon^2/2} \ .$$

Thus it suffices to have

$$\frac{\ln\left|\mathcal{H}_{k(n)}\right|}{n} \to 0, \text{ as } n \to \infty \ .$$

Another way to write this is using little-oh notation. We say that a sequence $a_n = o(b_n)$ if $\frac{a_n}{b_n} \to 0$ as $n \to \infty$. So, the above is equivalent to $\ln\left|\mathcal{H}_{k(n)}\right| = o(n)$. Under this condition,

$$\forall \epsilon > 0, \ \exists N_\epsilon, \ \forall n \geq N_\epsilon, \ \ \ln\left|\mathcal{H}_{k(n)}\right| - \frac{n\epsilon^2}{2} \leq -\frac{n\epsilon^2}{4} \ . \tag{2}$$

Then for all $\epsilon > 0$,

$$\sum_{n \geq 1} \Pr\left(R(\widehat{h}_n) - R^*_{\mathcal{H}_{k(n)}} \geq \epsilon\right) \leq N_\epsilon + \sum_{n \geq N_\epsilon} 2e^{-n\epsilon^2/4} < \infty \ .$$

The summation on the right is simply a converging geometric series. By Corollary 1 we have established

Theorem 2. *Let $\{\mathcal{H}_k\}$ satisfy the UAP with $\left|\mathcal{H}_{k(n)}\right| < \infty$. Let $k(n)$ be an integer sequence such that as $n \to \infty$, $k(n) \to \infty$ and $\ln\left|\mathcal{H}_{k(n)}\right| = o(n)$. Then $R(\widehat{h}_n) \to R^*$ almost surely, i.e., $\widehat{h}_n$ is strongly universally consistent.*

Corollary 2. *Let $\{\mathcal{H}_k\} = \{$the set of histogram classifiers with side length $1/k\}$, $\mathcal{X} = [0,1]^d$. Let $k(n) \to \infty$ such that $k^d = o(n)$. Then $\widehat{h}_n$ is strongly universally consistent.*

Proof. We previous saw in Section 2 that histograms have the UAP. Also, $\left|\mathcal{H}_{k(n)}\right| = 2^{k^d}$, so the corollary follows from Theorem 2. □

In Corollary 2, we require $\frac{k^d}{n} \to 0$. Note that n/k^d is the number of samples per cell. Therefore the number of samples per cell must go to infinity; this seems like a reasonable condition for strong consistency.

Now, let's consider the case where the VC dimension of $\mathcal{H}_k$ is finite. From our discussion of VC theory, we have that for all $\epsilon > 0$,

$$\Pr\left(R(\widehat{h}_n) - R^*_{\mathcal{H}_{k(n)}} \geq \epsilon\right) \leq 8S_{\mathcal{H}_{k(n)}}(n)e^{-n\epsilon^2/128} \ .$$

We also know from Sauer's lemma

$$S_{\mathcal{H}}(n) \leq \left(\frac{ne}{V}\right)^V, \ n \geq V,$$

where V is the VC dimension of $\mathcal{H}$. Using this, we deduce

$$\Pr\left(R(\widehat{h}_n) - R^*_{\mathcal{H}_{k(n)}} \geq \epsilon\right) \leq 8S_{\mathcal{H}_{k(n)}}(n)e^{-n\epsilon^2/128} \leq Cn^{V_{k(n)}}e^{-n\epsilon^2/128} = C\exp\left(V_{k(n)}\ln n - n\frac{\epsilon^2}{128}\right) \ .$$

C is simply a constant that does not depend on n. If we choose $k(n)$ such that $V_{k(n)} = o\left(\frac{n}{\ln n}\right)$, then we obtain the following inequality, which is similar to the one in Eqn. (2):

$$\forall\epsilon > 0, \ \exists N, \ \forall n \geq N, \ V_{k(n)}\ln n - \frac{n\epsilon^2}{128} \leq -\frac{n\epsilon^2}{256} \ .$$

Therefore

$$\sum_{n\geq 1}\Pr\left(R(\widehat{h}_n) - R^*_{\mathcal{H}_{k(n)}} \geq \epsilon\right) \leq N + \sum_{n\geq N} Ce^{-n\epsilon^2/256} < \infty \ .$$

We have established

Theorem 3. *Let $\{\mathcal{H}_k\}$ satisfy the UAP, and let $V_k < \infty$ denote the VC dimension of $\mathcal{H}_k$. Let $k(n)$ be an integer-valued sequence such that as $n \to \infty$, $k(n) \to \infty$ and $V_{k(n)} = o(\frac{n}{\ln n})$. Then $R(\widehat{h}_n) \to R^*$ almost surely, i.e., $\widehat{h}_n$ is strongly universally consistent.*

5 Rates of Convergence

One could ask whether there is a universal rate of convergence. Put more formally, is there an $\widehat{h}_n$ such that $\mathbb{E}\left[R(\widehat{h}_n) - R^*\right]$ converges to 0 at a fixed rate, for any joint distribution P_{XY}? It turns out that the answer is no. A theorem from [2] makes this concrete.

Theorem 4. *Let $\{a_n\}$ with $a_n \searrow 0$ such that $\frac{1}{16} \geq a_1 \geq a_2 \geq \ldots$. For any $\widehat{h}_n$, there exists a joint distribution P_{XY} such that $R^* = 0$ and $\mathbb{E}R(\widehat{h}_n) \geq a_n$.*

Proof. See Chapter 7 of [1]. □

Since the sequence $\{a_n\}$ is arbitrary, we can also choose how slowly it converges to 0, thus showing that there will be no universal rate of convergence. In order to establish rates of convergence, we will therefore have to place restrictions on P_{XY}. Some possible reasonable restrictions on P_{XY} include:

- $P_{X|Y=y}$ is a continuous random variable.
- $\eta(x)$ is "smooth".
- there exists a t_0 such that $P_X\left(\left|\eta(X) - \frac{1}{2}\right| \leq t_0\right) = 0$.
- the Bayes decision boundary is "smooth".

Below we consider one particular distributional assumption under which we can establish a rate of convergence.

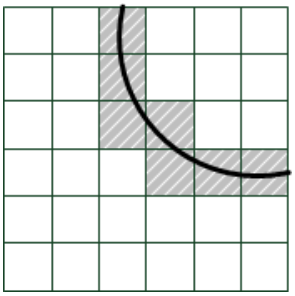

Figure 3: Illustration of Bayes decision boundary (BDB).

6 Box-Counting Class

Let $\mathcal{X} = [0,1]^d$. For $m \geq 1$, define $\mathcal{P}_m$ to be the partition of $\mathcal{X}$ into cubes of side length $1/m$. Note that $|\mathcal{P}_m| = m^d$.

Definition 3. *The box counting class $\mathcal{B}$ is the set of all P_{XY} such that*

(A) *P_X has a bounded density.*

(B) *There exists a constant C such that $\forall m \geq 1$, the Bayes decision boundary (BDB) $\{x : \eta(x) = 1/2\}$ passes through at most Cm^{d-1} elements of $\mathcal{P}_m$ (see Fig. 3).*

This essentially says that the Bayes decision boundary (BDB) has dimension $d-1$. Look up the term "box-counting dimension" for more.

We have the following rate of convergence result.

Theorem 5. *Let $\mathcal{H}_k = \{$histogram classifiers based on $\mathcal{P}_k\}$, assume $P_{XY} \in \mathcal{B}$, and let $\widehat{h}_n$ be a sieve estimator with and $k \sim n^{\frac{1}{d+2}}$. Then*

$$\mathbb{E}\left[R(\widehat{h}_n)\right] - R^* = O(n^{-\frac{1}{d+2}}).$$

Proof. Recall the decomposition into estimation and approximation errors:

$$\mathbb{E}\left[R(\widehat{h}_n) - R^*\right] = \mathbb{E}\left[R(\widehat{h}_n) - R^*_{\mathcal{H}_k}\right] + R^*_{\mathcal{H}_k} - R^*.$$

To bound the estimation error, denote $\widehat{\Delta}_n := R(\widehat{h}_n) - R^*_{\mathcal{H}_k}$. By the law of total expectation,

$$\mathbb{E}\left[\widehat{\Delta}_n\right] = \underbrace{Pr(\widehat{\Delta}_n \geq \epsilon)}_{\leq \delta} \underbrace{\mathbb{E}\left[\widehat{\Delta}_n \mid \widehat{\Delta}_n \geq \epsilon\right]}_{\leq 1 \text{ (trivial since } R \leq 1)} + \underbrace{Pr(\widehat{\Delta}_n < \epsilon)}_{\leq 1} \underbrace{\mathbb{E}\left[\widehat{\Delta}_n \mid \widehat{\Delta}_n < \epsilon\right]}_{\leq \epsilon}.$$

Choosing $\epsilon = \sqrt{\frac{2[k^d \ln 2 + \ln(2/\delta)]}{n}}$ and $\delta = 1/n$, we have $\mathbb{E}\left[R(\widehat{h}_n)\right] - R^*_{\mathcal{H}_k} = O\left(\sqrt{\frac{k^d}{n}}\right)$.

To bound the approximation error, let $h^* =$ Bayes classifier and $h^*_k =$ best classifier in $\mathcal{H}_k$. Observe

$$\begin{aligned} R(h^*_k) - R(h^*) &= 2\mathbb{E}_X\left[\left|\eta(x) - \frac{1}{2}\right| \mathbf{1}_{\{h^*_k(X) \neq h^*(X)\}}\right] \\ &\leq \mathbb{E}_X\left[\mathbf{1}_{\{h^*_k(X) \neq h^*(X)\}}\right] \quad \left(\text{by } \left|\eta(x) - \frac{1}{2}\right| \leq \frac{1}{2}\right) \\ &= P_X\left(h^*_k(X) \neq h^*(X)\right). \end{aligned} \tag{3}$$

Let $\Delta := \{x : h_k^*(x) \neq h^*(x)\}$, and let f be the density of P_X. Then

$$\begin{aligned} P_X(\Delta) &= \int_\Delta f(x)dx \\ &\leq \|f\|_\infty \cdot \lambda(\Delta), \end{aligned}$$

where λ denotes volume/Lebesgue measure. Now classification errors occur only on cells intersecting the Bayes decision boundary (see shaded cells in Fig. 3), and so $\lambda(\Delta) \leq Ck^{d-1}/k^d = O(\frac{1}{k})$. Setting $\frac{1}{k} = \sqrt{\frac{k^d}{n}}$, we see that $k \sim n^{\frac{1}{d+2}}$ gives the best rate for histograms. □

Histograms do not achieve the best rate among all discrimination rules for the box-counting class; this best rate is $\mathbb{E}\left[R(\widehat{h}_n)\right] - R^* = O(n^{-\frac{1}{d}})$ [4]. In the next set of notes we'll examine a discrimination rule that achieves this rate to within a logarithmic factor.

Exercises

1. With $\mathcal{X} = \mathbb{R}^d$ let

   $$\mathcal{H}_k = \{h(x) = \mathbf{1}_{\{f(x)\geq 0\}} \,:\, f \text{ is a polynomial of degree at most } k\}.$$

   Let $\mathcal{P}$ be the set of all joint distributions P_{XY} such that

   $$\inf_{h\in\mathcal{H}_k} R(h) \to R^*$$

   as $k \to \infty$. Determine an explicit sufficient condition on $k(n)$ for the sieve estimator to be consistent for all $P_{XY} \in \mathcal{P}$.

2. Up to this point in our discussion of empirical risk minimization, we have always assumed that an empirical risk minimizer exists. However, it could be that the infimum of the empirical risk is not attained by any classifier. In this case, we can still have a consistent sieve estimator based on an approximate empirical risk minimizer. Thus, let τ_k be a sequence of positive real numbers decreasing to zero. Define $\widehat{h}_{n,k}$ to be any classifier

   $$\widehat{h}_{n,k} \in \{h \in \mathcal{H}_k \,:\, \widehat{R}_n(h) \leq \inf_{h'\in\mathcal{H}_k} \widehat{R}_n(h') + \tau_k\}.$$

   Now consider the discrimination rule $\widehat{h}_n := \widehat{h}_{n,k(n)}$. Show that Theorems 2 and 3 still hold for this discrimination rule. State any additional assumptions on the rate of convergence of τ_k that may be needed.

3. Assume $\mathcal{X} \subseteq \mathbb{R}^d$. A function $f : \mathcal{X} \to \mathbb{R}$ is said to be Lipschitz continuous if there exists a constant $L > 0$ such that for all $x, x' \in \mathcal{X}, |f(x) - f(x')| \leq L\|x - x'\|$. Show that if one coordinate of the Bayes decision boundary is a Lipschitz continuous function of the others, then **(B)** holds in the definition of the box-counting class. An example of the stated condition is when the Bayes decision boundary is the graph of, say, $x_d = f(x_1, \ldots, x_{d-1})$, where f is Lipschitz.

4. Establish the following partial converse to Lemma 2: If $|Z_n - Z|$ is bounded, then $Z_n \xrightarrow{i.p.} Z$ implies $\mathbb{E}\{|Z_n - Z|\} \to 0$. Therefore, in the context of classification, $\mathbb{E}R(\widehat{h}_n) \to R^*$ is equivalent to weak consistency.

References

[1] L. Devroye, L. Györfi and G. Lugosi, *A Probabilistic Theory of Pattern Recognition*, Springer, 1996.

[2] L. Devroye, "Any discrimination rule can have an arbitrarily bad probability of error for finite sample size," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 4, pp. 154-157, 1982.

[3] U. Grenander, *Abstract Inference*, Wiley, 1981.

[4] C. Scott and R. Nowak, "Minimax-Optimal Classification with Dyadic Decision Trees," *IEEE Trans. Inform. Theory*, vol. 52, pp. 1335-1353, 2006.